IN THE UNITED STATES DISTRICT COURT
FOR THE WESTERN DISTRICT OF VIRGINIA
DANVILLE DIVISION

UNITED STATES OF AMERICA	)	Criminal Action No. 4:23-cr-00006
	)	
v.	)	**MEMORANDUM OPINION**
	)	
ISAAC JEROME GRAHAM,	)	By: Hon. Thomas T. Cullen
	)	United States District Judge
Defendant.	)	

On December 28, 2022, a masked man robbed a convenience store in Danville, Virginia. He fired a round from a handgun into the floor during the robbery. Police later arrested Defendant Isaac Jerome Graham, and, after he confessed to committing the robbery, recovered a Glock 19 pistol in his car. Police test-fired the Glock and sent the resulting two shell casings to the Virginia Department of Forensic Science ("DFS") to compare the test-fired casings with a shell casing found at the convenience store. Applying her department's toolmark analysis methodology and years of experience, DFS Forensic Scientist Laura Hollenbeck ("Hollenbeck") determined that the pistol recovered from Graham's car was used to fire the bullet into the convenience-store floor. The government intends to call Hollenbeck as an expert to testify to that effect at Graham's trial, scheduled to begin on April 15, 2024.

The matter is before the court on Graham's motion to exclude, or alternatively to curtail, Hollenbeck's testimony under *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 509 U.S. 579 (1993), and Federal Rule of Evidence 702. (Def.'s Mot. Exclude [ECF No. 68].) The court conducted a *Daubert* hearing on November 30, 2023, during which Hollenbeck testified about

her comparison and applied methodology. Following that hearing, Graham filed a renewed motion to exclude; the motion is now ripe for disposition.[1]

At bottom, Graham argues that the field of firearm and toolmark analysis is inherently flawed, that Hollenbeck's opinion, which is based on her application of the prevailing methodology in that field, is unreliable under *Daubert,* and, given recent scrutiny about the discipline, that Hollenbeck's testimony should be excluded. The court disagrees and will deny Graham's motion to exclude Hollenbeck's opinion. But based on Rule 702's recent amendments, the court will order that Hollenbeck's testimony conform with the U.S. Department of Justice's Uniform Language for Testimony of Reports for the Forensic Firearms/Toolmarks Discipline.

I. STANDARD OF REVIEW

Federal Rule of Evidence 702 governs the admissibility of expert-witness testimony, along with the Supreme Court's decisions in *Daubert* and *Kumho Tire Co., Ltd. v. Carmichael*, 526 U.S. 137 (1999). Collectively, these impose on the court a gatekeeping role in which it must ensure that proffered "expert evidence is sufficiently relevant and reliable *when it is submitted to the jury*." *Nease v. Ford Motor Co.*, 848 F.3d 219, 231 (4th Cir. 2017) (emphasis in original). Rule 702 provides that a qualified expert's opinion is admissible if:

> (a) the expert's scientific, technical, or other specialized knowledge will help the trier of fact to understand the evidence or to determine a fact in issue;

[1] Graham previously moved for the court to compel additional disclosures related to the methodology behind Hollenbeck's opinions under Federal Rule of Criminal Procedure 16(a)(1)(G), or, alternatively, to exclude or limit her testimony. (ECF No. 45.) Construing that motion as *Daubert* challenge, the court found it efficient to conduct a *Daubert* hearing. (Order, Nov. 1, 2023 [ECF No. 47].) Graham filed the instant motion on December 18, 2023. Neither party requests another hearing and much of Graham's argument is one that has been advanced across numerous federal courts; accordingly, the court dispenses with additional oral argument because it would not aid in the decisional process.

> (b) the testimony is based on sufficient facts or data;
> (c) the testimony is the product of reliable principles and methods; and
> (d) the expert's opinion reflects a reliable application of the principles and methods to the facts of the case.

Fed. R. Evid. 702. Ultimately, the court's objective "is to make certain that an expert, whether basing testimony upon professional studies or personal experience, employs in the courtroom the same level of intellectual rigor that characterizes the practice of an expert in the relevant field." *Kumho Tire*, 526 U.S. at 152.

In this gatekeeping role, the court must balance "two guiding, and sometimes competing principles: Rule 702 was intended to liberalize the introduction of relevant expert evidence[,] and expert witnesses have the potential to be both powerful and quite misleading." *Hickerson v. Yamaha Motor Corp.*, 882 F.3d 476, 481 (4th Cir. 2018) (cleaned up). *Daubert* instructed, consistent with that balance, that "[v]igorous cross-examination, presentation of contrary evidence, and careful instruction on the burden of proof are the traditional and appropriate means of attacking shaky but admissible evidence." *Daubert*, 509 U.S. at 596; *see Kovari v. Brevard Extraditions, LLC*, 461 F. Supp. 3d 353, 369 (W.D. Va. 2020). But before the court may admit even "shaky" expert testimony, the proponent of that evidence must establish by a preponderance of evidence that the testimony is admissible. *See* Fed. R. Evid. 702. The court has "considerable leeway" in determining whether the proponent has carried his burden, *see Hickerson*, 882 F.3d at 480, but it may not "abandon the gatekeeping function." *Nease v. Ford Motor Co.*, 848 F.3d 219, 230 (4th Cir. 2017) (quoting *Kumho Tire*, 526 U.S. at 158–59 (Scalia, J., concurring)).

II. ANALYSIS

Before opening the gate to Hollenbeck's testimony, the government must show that she is qualified and that her testimony is relevant and reliable. *United States v. Peterson*, No. 1:19-cr-00054, 2020 WL 5039504, at *4 (W.D. Va. Aug. 26, 2020). Graham does not meaningfully challenge Rule 702's threshold consideration, and the court finds Hollenbeck's years of experience qualifies her to testify as an expert in the field of firearm-toolmark analysis.[2] Graham also does not contest that Hollenbeck's testimony is relevant. It is beyond dispute that her testimony "will help the trier of fact to understand the evidence or to determine a fact in issue"—specifically, whether the firearm recovered from Graham was discharged during the robbery. Fed. R. Evid. 702(a); (*see* Gov't Resp. at 13.)

Instead, Graham argues that the field of firearm-toolmark analysis and identification—and Hollenbeck's application of that methodology—are unreliable. (*See generally* Def.'s Mot. Exclude.)

A. Background on Firearm-Toolmark Analysis

Before analyzing Hollenbeck's testimony, the court takes a necessary detour to provide some background on the discipline at the heart of Graham's motion and how courts have

[2] Both at the hearing and on brief, Graham implied Hollenbeck is not qualified because of what he thinks are deficiencies in her formal education. (*See* Def.'s Mot. Exclude at 2.) The court construes this as zealous advocacy by counsel but not a bona fide challenge to Hollenbeck's qualifications as an expert witness. Indeed, Rule 702 does not require an expert witness to have a formal education, much less thrive in her schooling. *See* Fed. R. Evid. 702 ("A witness who is qualified as an expert by knowledge, skill, experience, training, *or* education . . . .") (emphasis added). Hollenbeck's years of experience as a forensic toolmark examiner qualify her "to testify about a specialized area outside the knowledge of the average juror." *United States v. Diaz*, No. 22-4277, 2023 WL 6366689, at *2 (4th Cir. Sept. 29, 2023); (*see also* ECF No. 55-2 (listing 28 state court cases in which Hollenbeck testified as an expert).)

previously treated testimony based on firearm-toolmark identification.[3]

1. Firearm-Toolmark Analysis's Methodology

The assumption at the heart of firearm-toolmark analysis is common among forensic sciences: no two firearms are exactly alike. When a gun is assembled in the factory, the firearm manufacturing tools "wear during their use and change microscopically." Jaimie A. Smith, *Beretta Barrell Fired Bullet Validation Study*, 66 J. Forensic Scis. 547, 547 (2020) (available at ECF No. 71-5) [hereinafter Smith study]; *see* Nat'l Rsch. Council, *Strengthening Forensic Science in the United States: A Path Forward* 150 (2009), https://www.ojp.gov/pdffiles1/nij/grants/228091.pdf (last visited February 20, 2024) [hereinafter 2009 NRC Report]. Because the tools that create firearms change microscopically with each metal-on-metal clash, they impart "random imperfections" on a gun's firing components, including its barrel and firing pin. Smith study at 547. Those microscopic imperfections essentially give each firearm its own uniquely identifiable fingerprint, which it impresses upon ammunition each time it is discharged. (Hr'g Tr. 17:4–5 [ECF No. 61]); *see United States v. Pete*, No. 3:22-cr-00048, 2023 WL 4928523, at *1 (N.D. Fla. July 21, 2023). Those impressions are referred to as "toolmarks."

Some, but not all, toolmarks on a bullet or casing are unique to a single gun. Toolmarks fall into one of three categories, ranging from more to less common: (1) class characteristics;

[3] In arguing that the court should reject Hollenbeck's testimony, Graham urges the court to conduct its own analysis and not simply rely on the fact that no court has entirely excluded the kind of testimony Hollenbeck intends to give. (Def.'s Mot. Exclude at 5–9.) Graham's point is well-taken. Although other courts' near-universal acceptance of testimony similar to Hollenbeck's may bear on certain *Daubert* factors, there is no *per se* rule mandating admission. The court therefore undertakes its own independent analysis of the methodology's reliability.

(2) subclass characteristics; and (3) individual characteristics. *See United States v. Harris*, 502 F. Supp. 3d 28, 34–35 (D.D.C. 2020). A class characteristic is an intentional design feature that "will be present in all weapons of the same make and model." *United States v. Willock*, 696 F. Supp. 2d 536, 558 (D. Md. 2010) (citation omitted). Subclass characteristics are unintentional marks that will be present in a batch of guns "manufactured using the same equipment around the same time." *Pete*, 2023 WL 4928523, at *1. Individual characteristics are those "unique, microscopic, random imperfections in the barrel or firing mechanism created by the manufacturing process and/or damage to the gun post-manufacture, such as striated and/or impressed marks, unique to a single gun." *Harris*, 502 F. Supp. 3d at 34–35.

Using a comparison microscope, firearm-toolmark examiners compare the marks on recovered ammunition to those on test-fired ammunition "to determine whether ammunition is or is not associated with a specific firearm." President's Council of Advisors on Sci. & Tech., *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods* 104 (2016), https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf (last visited February 20, 2024) [hereinafter PCAST Report]. In conducting these comparisons, examiners typically follow the theory of identification propounded by the Association of Firearm and Tool Mark Examiners ("AFTE"). *See United States v. Ashburn*, 88 F. Supp. 3d 239, 246 (E.D.N.Y. 2015) (noting the AFTE theory is "the field's established standard"). When properly conducted, the AFTE comparison method "permits an examiner to conclude that two bullets or two cartridges are of common origin . . . when the microscopic surface contours of their toolmarks are in 'sufficient agreement.'" *United States v. Otero*, 849 F. Supp. 2d 425, 431 (D.N.J. 2012). Sufficient

agreement exists between two samples if "the agreement of individual characteristics is of a quantity and quality that the likelihood another tool could have made the mark is so remote as to be considered a practical impossibility." AFTE, *AFTE Theory of Identification as it Relates to Toolmarks*, https://afte.org/about-us/what-is-afte/afte-theory-of-identification (last visited February 20, 2024) [hereinafter AFTE Theory of Identification].

The examiner's threshold question under the AFTE methodology is whether two samples share class characteristics; if they do not, the samples objectively could not have come from the same gun.[4] *See United States v. Briscoe*, No. 20-cr-1777, 2023 WL 8096886, at *5 (D.N.M. Nov. 21, 2023); *United States v. Felix*, No. 20-cr-0002, 2022 WL 17250458, at *4 (D.V.I. Nov. 28, 2022) (noting this part of the process is objective). If the class characteristics are the same, an examiner then compares the subclass and individual characteristics. *See Briscoe*, 2023 WL 8096886, at *5. Following that comparison, an examiner must come to one of four conclusions:

1. **Identification:** Agreement of all discernible class characteristics and sufficient agreement of a combination of individual characteristics.

2. **Elimination:** Significant disagreement of discernible class characteristics and/or individual characteristics.

3. **Inconclusive:** Agreement of all discernible class characteristics [A] and some agreement of individual characteristics, but

[4] The first comparison may be done using the Bureau of Alcohol, Tobacco, Firearms and Explosive's ("ATF") National Integrated Ballistic Information Network ("NIBIN"). (Hr'g Tr. 26:6.) NIBIN is a database of "three dimensional digital ballistic images of spent shell casings recovered from crime scenes and from crime gun test-fires that can automatically generate a list of potential matches, purportedly with a very high level of accuracy." *United States v. Hunt*, 63 F.4th 1229, 1239 (10th Cir. 2023) (internal quotations omitted); (*see also* Hr'g Tr. 24:24–25:11.) In essence, NIBIN runs an algorithm that compares the samples at a macro level and identifies whether there are potential associations, at which point an examiner must confirm the virtual comparison with her microscopic comparison. (Hr'g Tr. 26:7–19); *see also Hunt*, 63 F.4th at 1239 (describing NIBIN as a tool that provides leads that an examiner must confirm).

> insufficient for an identification . . . [B] . . . without agreement or disagreement of individual characteristics due to an absence, insufficiency, or lack of reproducibility . . . [C] . . . and disagreement of individual characteristics, but insufficient for an elimination.

4. **Unsuitable:** Unsuitable for examination.

AFTE, *Technical Procedures Manual* 110 (2015) (available at ECF No. 71-1) (cleaned up) [hereinafter AFTE Manual]. Individual-characteristic analysis is "subjective in nature" but is "founded on scientific principles and based on the examiner's training and experience." AFTE Theory of Identification.

Hollenbeck has been a member of the AFTE since 2013 and has been certified in firearms identification through the organization since 2020. (*See* Hollenbeck Curriculum Vitae [ECF No. 55-1].) Her AFTE training notwithstanding, Hollenbeck testified that she follows the DFS's Firearm/Toolmark Procedures manual when she conducts a toolmark comparison. (Hr'g Tr. 15:23–16:1.) The DFS's manual "is based heavily on the AFTE procedures manual[,]" and the DFS's procedure shares the same basic theory of identification as the AFTE's. (*Id.* 15:21–22, 16:14–15); *see also* DFS, *Firearm/Toolmark Procedures Manual* 5 (2023) (available at ECF No. 71-2) [hereinafter DFS Manual]. The court's review of both manuals confirms that characterization. Hollenbeck's testimony—and Graham's motion—treat the AFTE's methodology and DFS's methodology as one and the same for *Daubert* purposes, and the court agrees that any distinctions are negligible and immaterial.[5]

[5] If anything, DFS's manual appears to impose more requirements on its examiners that, in the court's mind, only increase the method's reliability. For example, the AFTE "strongly recommend[s]" documentation of "observations that support a reported conclusion," AFTE Manual at 110–11, while the DFS requires such documentation, *see, e.g.*, DFS Manual at 23 (instructing that the marks that support an identification or elimination must "be photographed and/or described in examination documentation").

2. Trends in Admission

Courts have long permitted expert testimony tying a fired bullet to a specific gun based on marks on the recovered ammunition. *See, e.g.*, *Massachusetts v. Best*, 62 N.E. 748, 750 (Mass. 1902) (Holmes, C.J.). For much of the twentieth century, and into the early 2000s, firearm examiners have routinely been allowed to testify, based on toolmark analysis, that ammunition recovered from a crime scene was fired from a particular gun. *See United States v. Romero-Lobato*, 379 F. Supp. 3d 1111, 1116 (D. Nev. 2019). Given that rote acceptance, courts were historically reluctant to challenge the discipline's reliability. *See United States v. Monteiro*, 407 F. Supp. 2d 351, 364 (D. Mass 2006); *see United States v. Green*, 405 F. Supp. 2d 104, 109 (D. Mass. 2005) (admitting limited toolmark testimony based on the court's "confidence that any other decision will be rejected by appellate courts, in light of precedents across the country"). Even so, "storm clouds . . . gather[ed]" as courts began to question its reliability in a post-*Daubert* and *Kumho Tire* evidentiary landscape. *Monteiro*, 407 F. Supp. 2d at 364.

In 2008 and 2009, the National Research Council ("NRC") published two reports that critiqued the theories underlying the field, challenged the reliability of firearm and toolmark identification generally, and undermined confidence in the AFTE theory of identification. *See generally* 2009 NRC Report (citing the 2008 report throughout). The reports critiqued the AFTE methodology specifically, chiding its "lack of a precisely defined process." *See id.* at 155. In 2016, the President's Council of Advisors on Science and Technology ("PCAST") published its own report, concluding that the discipline fell short of "foundational validity, because there [was] only a single appropriately designed study to measure validity and estimate reliability." PCAST Report at 112. Emboldened by these reports, litigants began to challenge

toolmark experts' testimony, and courts no longer "automatically accept expert testimony derived from the AFTE method." *Romero-Lobato*, 379 F. Supp. 3d at 1117. Although—to the court's knowledge—no court has ever entirely excluded the kind of testimony the government hopes to introduce here, many courts impose limits.[6] *See, e.g.*, *United States v. Davis*, No. 4:18-cr-00011, 2019 WL 4306971, at *4 (W.D. Va. Sept. 11, 2019).

Still, as the Tenth Circuit advised, "in light of the critiques expressed in the PCAST and NRC Reports, . . . courts should be cautious" in admitting firearm-toolmark testimony. *United States v. Hunt*, 63 F.4th 1229, 1244 (10th Cir. 2023). The amendments to Federal Rule of Evidence 702 instruct courts to take a similar level of caution in admitting "testimony of forensic experts . . . if the methodology is subjective and thus potentially subject to error." Fed. R. Evid. 702 advisory committee's note to 2023 amendments.

With that background, the court turns to the reliability of Hollenbeck's testimony.

B. Hollenbeck's Testimony is Reliable

"Reliability is a 'flexible' inquiry that focuses on 'the principles and methodology' employed by the expert." *Sardis v. Overhead Door Corp.*, 10 F.4th 268, 281 (4th Cir. 2021) (quoting *Daubert*, 509 U.S. at 594–95). Testimony is not reliable unless the "expert's opinion is based on scientific, technical, or other specialized knowledge and not on belief or

[6] Indeed, even the two cases upon which Graham relies most heavily—*United States v. Tibbs*, No. 2016-CF1-19431, 2019 WL 4359486 (D.C. Super. Ct. Sept. 5, 2019), and *United States v. Adams*, 444 F. Supp. 3d 1248 (D. Or. 2020), which are somewhat anomalous in the extent of their critique of the discipline—did not completely exclude testimony from the government's toolmark expert. *Adams*, 444 F. Supp. 3d at 1267 (allowing the expert to testify to observational evidence made during his toolmark comparison); *Tibbs*, 2019 WL 4359486, at *1 (permitting the expert to testify that, "based on his examination of the evidence and the consistency of the class characteristics and microscopic toolmarks, the firearm cannot be excluded as the source of the casing").

speculation." *Id.* (cleaned up). To guide that inquiry, *Daubert* provided five factors that a court may consider:

> (1) whether the particular scientific theory "can be (and has been) tested"; (2) whether the theory "has been subjected to peer review and publication"; (3) the "known or potential rate of error"; (4) the "existence and maintenance of standards controlling the technique's operation"; and (5) whether the technique has achieved "general acceptance" in the relevant scientific or expert community.

United States v. Crisp, 324 F.3d 261, 266 (4th Cir. 2003) (quoting *Daubert*, 509 U.S. at 593–94). After finding the underlying methodology reliable, the court must ensure "the expert's opinion reflects a reliable application" of that methodology. Fed. R. Evid. 702(d); *see Acosta v. Vinoskey*, 310 F. Supp. 3d 662, 669 (W.D. Va. 2018).

1. Hollenbeck's Methodology Passes Muster Under *Daubert*

Graham's motion focuses largely on his claim that the reliability problems with Hollenbeck's testimony are those associated with firearm-toolmark analysis writ large.[7] (*See* Def.'s Mot. Exclude at 1.) The court agrees that toolmark analysis is not the most precise science, but—as discussed below—*Daubert*'s five factors demonstrate that Hollenbeck's methodology is sufficiently reliable to be admissible.

i. Testability

A theory's testability is "a key question" in a court's reliability analysis. *Daubert*, 509 U.S. at 593. This factor "concerns 'whether the expert's theory can be challenged in some

[7] At the outset, the court notes that most of Graham's *Daubert* analysis mirrors the court's opinion in *Tibbs*, 2019 WL 4359486. While that court's reasoning is undoubtedly thorough (and at times persuasive), the court does not agree with its view of gatekeeping. "In fulfilling its gatekeeping function, a district court must conduct a preliminary assessment to determine whether the methodology underlying the expert witness' testimony is valid." *Bresler v. Wilmington Tr. Co.*, 855 F.3d 178, 195 (4th Cir. 2017) (cleaned up). In the court's view, the *Tibbs* court's assessment, which dives deep into statistical and scientific theory, exceeds *Daubert*'s charge.

objective sense, or whether it is instead simply a subjective, conclusory approach that cannot be reasonably assessed for reliability.'" *Harris*, 502 F. Supp. 3d at 37 (quoting Fed. R. Evid. 702 advisory committee's note to 2000 amendments). Graham agrees with the general premise "that examiners can be given a casing fired from a known firearm to determine whether they have reached the correct conclusion." (Def.'s Mot. Exclude at 9.) But Graham asserts that "the method is precisely not testable" because "we cannot know why the examiner reached the conclusion." (*Id.*)

In contending that Hollenbeck's method is not testable, Graham echoes the court's analysis in *Adams*, 444 F. Supp. 3d 1248, which appears to stand alone in its conclusion that the AFTE method is not testable. There, the court summarized a toolmark expert's opinion as "applying his training and experience to make a subjective conclusion about what he sees before him." *Id.* at 1263–64. That subjective step, according to the *Adams* court, rendered the "AFTE method . . . not testable . . . because it cannot be explained in a way that would allow an *uninitiated* person to perform the same test in the same way that [the expert] did." *Id.* at 1264 (emphasis added). *Adams*, however, undermines that reasoning in the prior paragraph by discussing how a cancer researcher's opinion could be tested because "another cancer researcher" could perform the same test as the first researcher for diagnostic purposes. *Adams*, 444 F. Supp. 3d at 1263. To reframe *Adams*'s reasoning, the court would not hold a seasoned oncologist's tried-and-true diagnostic methodology untestable and unreliable for the sole reason that a layperson could not reach the same diagnosis after looking at the same scans.

Notably, another court in the same district as *Adams* recently disagreed with its analysis. *See United States v. Rhodes*, No. 3:19-cr-00333, 2023 WL 196174, at *2 (D. Or. Jan. 17, 2023).

The *Rhodes* court noted that the *Adams* court's "focus on an uninitiated person" was misplaced because *Daubert* testability requires that "someone else using the same data and methods be able to replicate the results." *Id.* at *3 (cleaned up) (quoting *City of Pomona v. SQM N. Am. Corp.*, 750 F.3d 1036, 1047 (9th Cir. 2014)). That "someone else" refers to an expert, not a layperson as the *Adams* court suggested and Graham argues. *Id. Adams* does not persuade the court.

The court is further taken by the fact that the PCAST Report—notable for its criticisms of firearm-toolmark analysis—not only refutes the idea that the methodology is not testable, but also prescribes the kind of test that can establish scientific validity for toolmark analysis and other "subjective" fields. *See* PCAST Report at 106. The notion that an analysis is testable despite having a subjective component is not surprising, as an expert's qualitative final judgment is not a smoking gun that precludes a finding of testability or reliability. *See Harris*, 502 F. Supp. 3d at 37; *United States v. Aman*, 748 F. Supp. 2d 531, 541, 541 n.15 (E.D. Va. 2010) (noting that subjective "[j]udgment is, and must be, ubiquitous in science" and collecting cases that admit expert testimony from disciplines that rely on qualitative judgments).

In the instant matter, the government has carried its burden in showing Hollenbeck's methodology is testable. The government provided several studies, and cited even more, that confirm that firearm-toolmark analysis can be—and has been—extensively tested. (*See* Gov't Resp. Exs. C–G [ECF Nos. 71-3 to 71-7]); *contra Crisp*, 324 F.3d at 274 (Michael, J., dissenting) (claiming the government did not carry its burden on this factor in part due to its failure to "introduce evidence of studies or testing that would show that fingerprint identification is based on reliable principles and methods"). Also persuasive is the fact that Hollenbeck takes annual proficiency tests. *See Rhodes*, 2023 WL 196174, at *2–3. And the court cannot ignore

that the overwhelming majority of courts have found the AFTE's methodology testable. *See United States v. Chavez*, No. 15-cr-00285, 2021 WL 5882466, at *2 (N.D. Cal. Dec. 13, 2021) ("Courts across this country nearly uniformly conclude that AFTE methodology can, and has been tested.") (collecting cases). Indeed, even the two courts that are paralleled perhaps only by *Adams* in their attack of the AFTE methodology find that this factor weighs in favor of reliability. *See Tibbs*, 2019 WL 4359486, at *7; *Abruquah v. Maryland*, 296 A.3d 961, 988 (Md. 2023) ("[I]t is undisputed that firearms identification can be tested.").

Because the AFTE methodology is testable, this factor weighs in favor of reliability.

ii. Peer Review and Publication

The next *Daubert* factor asks if "the theory or technique has been subjected to peer review and publication," an important step in determining reliability because such public exposure "increases the likelihood that substantive flaws in methodology will be detected." *Daubert*, 509 U.S. at 593. Graham argues that the AFTE methodology has never undergone meaningful peer review because the AFTE Journal—the field's most prolific journal—is not double-blind or available to non-AFTE members and is shrouded with bias. (*See* Def.'s Mot. Exclude at 13–14.) The government counters that (1) the AFTE Journal's peer review is sufficiently rigorous, and (2) even if it were not, publication outside that single journal has subjected the AFTE methodology to "meaningful peer review." (*See* Gov't Resp. at 18–20.) Because the government is correct on the latter point, the court will not address its first point.[8]

[8] The court declines to decide whether the AFTE Journal provides so-called "meaningful peer review" and endorses the *Harris* court's query that "excluding certain journals from consideration based on the type of peer review the journal employs goes beyond a court's appropriate gatekeeping function under *Daubert*." *Harris*, 502 F. Supp. 3d at 40. The *Tibbs* court found that the AFTE Journal's peer review process was not "meaningful" because it did not employ double-blind peer review, despite acknowledging that "neither *Daubert* . . . nor Rule

The fact that numerous studies have been published in journals that employ Graham's high standards for peer review confirms that this forensic discipline is routinely peer reviewed. The government identifies three articles detailing validation studies about the performance of examiners who use the AFTE methodology that were published in the Journal of Forensic Sciences. (Gov't Resp. Exs. E–G). According to the *Tibbs* court, studies "published in the Journal of Forensic Science[s] . . . have undergone meaningful peer review," because that journal is independent and undergoes double-blind peer review. 2019 WL 4359486, at *8 *see also Hunt*, 63 F.4th at 1248 (same); *Briscoe*, 2023 WL 8096886, at *8 (same). Accordingly, the court is satisfied that the AFTE methodology has been subjected to peer review.

Further supporting this point is the expert list to which the government points and that was relied on in *Harris*, 502 F. Supp. 3d 28 in 2020,[9] which identifies 47 studies related to firearms-examination research that were published in 11 peer-reviewed journals and not the AFTE Journal. (Gov't Resp. at 18 (citing *Harris*, No. 1:19-cr-00358, ECF No. 28-6 at 39–42.) The court has not independently examined each journal to assess its specific level of peer review, but notes that 15 of those studies were published in the Journal of Forensic Sciences, *Tibbs*'s gold-standard for "meaningful" peer review. (*Id.*) Graham also does not contest that these studies have subjected the AFTE methodology and theory of identification to sufficient peer review and publication scrutiny under *Daubert*.

---

702 mandate any specific type of peer review process." *See Tibbs*, 2019 WL 4359486, at *9. But as the *Harris* court explained, "there is far from consensus in the scientific community that double-blind peer review is the only meaningful kind of peer review." *Harris*, 502 F. Supp. 3d at 40. Where the scientific field cannot agree on what peer review is 'meaningful,' the court will not usurp its prerogative and do so itself. *Cf. Daubert*, 509 U.S. at 600–01 (Rehnquist, C.J., concurring in part and dissenting in part).

[9] According to the expert, his "declaration [was] written in response to" the *Tibbs* opinion. *Harris*, 502 F. Supp. 3d 28, ECF No. 28-6, at 2.

Because studies on the AFTE methodology have undergone peer-review that satisfies even Graham's high standard, this factor weighs in favor of reliability. *See Harris*, 502 F. Supp. 3d at 40; *see also Briscoe*, 2023 WL 8096886, at *8 (discounting the AFTE Journal, but finding the "AFTE Theory has been peer reviewed" on the back of three "peer reviewed studies on error rates and other topics in the field of toolmark analysis" that were published in the Journal of Forensic Sciences)

iii. Rate of Error

Next, the court is asked to "consider the known or potential rate of error" of a specific technique. *Daubert*, 509 U.S. at 594. As Hollenbeck testified, there is no error rate that applies to the field of toolmark identification at large, but the methodology's efficacy is measured with error rates from specific validation studies. (Hr'g Tr. 20:24–21:4.) Despite the numerous validation studies cited by the government that reveal a low error rate (*see* Gov't Resp. at 21–23), Graham argues that "[t]here is no known rate of error, and potential error rates vary wildly" (Def.'s Mot. Exclude at 15). According to Graham, the error rates cited by the government are invalid because of design flaws that plague each validation study. Specifically, Graham seizes on the fact that the studies do not mimic regular casework (in that participants know during the studies that a match exists within the set they are examining, whereas in the "field," it's possible, if not downright likely, that a match does not exist). He further points out that only "false positives" are counted as "errors" while inconclusive responses (where an examiner is unable to determine, one way or the other, if two samples are from the same

firearm) are not.[10] The court disagrees and finds the government has provided studies that comply even with Graham's and PCAST's design criticisms and still reveal a low potential error rate.

As an initial matter, a study's "error rate" that is only made up of "false positives" makes sense in this context. The court's gatekeeping role is meant to "protect the judicial process from" unreliable evidence that could impermissibly influence a jury. *Sardis*, 10 F.4th at 275. In the criminal context, the greatest risk that could arise from a court's failure to gatekeep is a false-positive identification that leads to a conviction. In the *Daubert* error-rate context, therefore, "the focal point of the inquiry should be on the rate of false positives, 'as this is the type of error that could lead to a conviction premised on faulty evidence.'" *Chavez*, 2021 WL 5882466, at *3 (quoting *Harris*, 502 F. Supp. 3d at 39); *Rhodes*, 2023 WL 196174, at *4 ("[W]hile an inconclusive result is an error insofar as it means the methodology did not produce an answer, it is not an error in the sense that it falsely attributes a cartridge or casing to the wrong firearm.").

In addition to chiding the treatment of inconclusive results, Graham argues that the court cannot trust the false-error rates provided in closed-set validation studies because they

[10] Graham also claims error rates are skewed because validation studies do not include as errors results from participants who drop out of a study before its completion and because study participants are typically volunteers. (*See* Def.'s Mot. Exclude at 18–19.) These arguments are almost entirely speculative. Hollenbeck confirmed at the hearing that an individual who is not confident in his identification *may* drop out of a study as a result, which could skew the study's error rate. (Hr'g Tr. 43:18–24.) Graham takes that possibility as an indictment on all validation studies. But Graham does not consider the inverse possibility; perhaps each study has hundreds more examiners ready to ace their identifications, but each has something come up that prevents her from completing the study. If Graham wants to persuasively advance a theory that toolmark examiners are so desperate to insulate their field from high error rates as to drop out of a study when they are not confident in their identification, he needs to rely on something more than rank speculation. *Cf. Long v. Hooks*, 947 F.3d 159, 183 (4th Cir. 2020) (Thacker, J., dissenting) ("[S]urely [counsel] is aware (or at least should be) that it is elemental that counsel's arguments are *not* evidence in a case. It is literally black letter law.").

do not mirror real-life casework. PCAST leveled the same criticism: the "'closed-set' design is simpler than the problem encountered in casework, because the correct answer is always present in the collection." PCAST Report at 108. On the other hand, "in an open-set study (as in casework), there is no guarantee that the correct source is present—and thus no guarantee that the closest match is correct."' *Id.* This criticism is fair;[11] closed-set studies may artificially deflate the false-positive error rate because "examiners can perform perfectly if they simply match each bullet to the standard that is *closest*." *Id.* (emphasis in original). Still, the PCAST Report praised one study—the Ames I study (*see* Gov't Resp. Ex. C [ECF No. 71-3])—as being "appropriately designed to test foundational validity and estimate reliability." PCAST Report at 111. The relevant false-positive error rate in that study was 1.01%.[12]

The government cites three more studies, published after the PCAST Report and designed like the Ames I study, that found even lower false-positive error rates: the Keisler, Smith, and Ames II studies.[13] (Gov't Resp. at 23.) Hollenbeck testified specifically that she had read and relied on the Smith study. (Hr'g Tr. 36:24.) The Smith study "was designed to answer some of the[] criticisms" from the PCAST Report and used "an 'open-set' design to help the discipline of firearm identification establish 'Foundational Validity' which is outlined

[11] The government challenges the PCAST Report's criticism, noting that "the PCAST Report's insistence on particular design features is both arbitrary and unsupported by the scientific literature." (Gov't Resp. at 22.) Nevertheless, in the court's mind, the report's criticism is reasonable.

[12] PCAST appears to have errantly listed this number as 1.5%. PCAST Report at 111; *see Tibbs*, 2019 WL 4359486, at *15 (citing a common criticism against the PCAST Report was that it "apparently miscounted or omitted data from several studies").

[13] Graham's motion noted that, "should the government actually offer any given study, [he would] respond after evaluating its relative strength." (Def.'s Mot. Exclude at 15.) Graham preemptively identified and challenged some of the studies cited by the government, but he did not file a reply brief to address others, so the court considers the non-challenged studies to be conceded as accurate.

in the PCAST Report." Smith study at 547. The false-positive rate in that study was 0.56%.[14] *Id.* at 551.

Graham challenges the Smith study's validity for several unpersuasive reasons, including that inconclusive results were not counted as errors and that the samples compared in that study were bullets rather than shell casings. (Def.'s Mot. Exclude at 19.) As discussed, the relevant error rate is that of false-positive identifications.[15] Next, because the same underlying AFTE theory of identification and methodology apply to both bullets and shell casings, the study is still relevant and persuasive. *See* PCAST Report at 104.

In sum, Graham's criticisms may have been well-founded years ago, but the government demonstrates that recent experiments ameliorated the major concerns surrounding validation studies. The open-set design studies suggest the potential error rate for an examiner like Hollenbeck to make a false identification is sufficiently low—about 1%. Further, as the *Chavez* court noted, that already low potential error rate becomes infinitesimal when the court considers that a toolmark examiner, like Hollenbeck, has her work verified by another examiner. 2021 WL 5882466, at *4. Applying *Chavez*'s formula, the chance of both examiners falsely identifying is approximately 0.01%. *See id.* at *4 n.2.

[14] The Smith study was published in the peer-reviewed Journal of Forensic Sciences. *See* Smith study at 547; *contra Tibbs*, 2019 WL 4359486, at *16 (discounting the Ames I and Keisler studies, which followed PCAST's design instructions, because they did not undergo "meaningful, independent peer review prior to publication").

[15] The same reason renders Graham's challenge that the Ames II study's repeatability and reproducibility rates suggest a high overall error rate unpersuasive. (*See* Def.'s Mot. Exclude at 20–21.) As the government points out, those concerns generally involved different classifications within the three inconclusive categories of that study; they did not result in an elevated false-positive error rate. (*See* Gov't Resp. at 27–29.) Because the false-positive rate remained low—indeed, well below 1%—throughout the Ames II study, the court finds the study relevant and persuasive.

Because the government has shown the potential error rate is low, this factor weighs in favor of reliability.

iv. Existence and Maintenance of Standards

The court next considers "the existence and maintenance of standards controlling the technique's operation." *Daubert*, 509 U.S. at 594. Graham claims this factor weighs against admissibility because AFTE methodology is circular and relies too heavily on an examiner's personal judgment. (Def.'s Mot. Exclude at 22.) The government counters that the methodology, despite its inescapable subjectivity, is "tightly controlled" by regular training and independent verification of examiner's conclusions. (Gov't Resp. at 29.)

Unlike the four other factors, courts appear split on this one in connection with toolmark analysis. Some courts agree with the government and find the factor weighs towards reliability. *See Otero*, 849 F. Supp. 2d at 435; *Rhodes*, 2023 WL 196174, at *5–6. For example, the *Rhodes* court found that specific procedures, training, regular proficiency testing, review by a second examiner, and the requirement that an examiner photograph and take extensive notes about his process outweighed the innate subjectivity in AFTE's methodology. 2023 WL 196174, at *5. Other courts, however, agree with Graham, finding those safeguards do not make up for the inherent subjectivity that drives each conclusion. *See Harris*, 502 F. Supp. 3d at 42; *Briscoe*, 2023 WL 8096886, at *11.

The court agrees with Graham on this factor. Though AFTE methodology is not entirely devoid of standards that help "maintain[] reliable results," it lacks the objective components for generating a conclusion that this factor contemplates. *Harris*, 502 F. Supp. 3d at 42. Like the expert in *Briscoe*, Hollenbeck's testimony indicated that her conclusion

"depend[ed] largely on a 'you know when you see it' methodology." *Briscoe*, 2023 WL 8096886, at *9; (Hr'g Tr. 70:20–24.) That sort of conclusion starkly contrasts with the very case to which *Daubert* cites in support of this factor's utility. *See Daubert*, 509 U.S. at 594 (citing *United States v. Williams*, 583 F.2d 1194, 1198 (2nd Cir. 1978)). *Williams* held that a professional organization's requirement that a set number of matches "be found before a positive identification can be made" is an indicia of the existence and maintenance of standards. *Williams*, 583 F.2d at 1198. The lack of objective standards governing the most important part of a toolmark examiner's process cuts against reliability.

In arguing against this conclusion, the government cites to *Crisp*, 324 F.3d at 268, a case that found latent-fingerprint analysis reliable despite its reliance on an examiner's subjective comparison. (*See* Gov't Resp. at 31.) But that case is distinguishable as it relates to this factor. In *Crisp*, the appellant, like Graham, argued that "fingerprint examiners operate without uniform, objective standards," and there was "no generally accepted standard regarding the number of points of identification necessary to make a positive identification." 324 F.3d at 268. While the Fourth Circuit rejected that argument and found the methodology reliable in part because it had some uniform standards that are also present in the field of firearm-toolmark analysis—*i.e.*, testing and proficiency requirements, "double checking"—it *also* found persuasive the fact that every identification relied on "a consistent 'points and characteristics'" approach to identification, despite the fact that there was not a uniform number of points necessary to make a positive identification field-wide. *Id.* at 269. In other words, a fingerprint examiner must be able to identify some number of points, even if that

number is not uniform in the field. Toolmark analysis, on the other hand, lacks even that level of objectivity. *See Harris*, 502 F. Supp. 3d at 41–42.

In sum, the unbridled subjectivity embedded in each toolmark examiner's conclusions cuts against reliability. But, like other courts have held, that cut is not fatal. *See Chavez*, 2021 WL 5882466, at *5. A methodology's subjectivity does not mean that it is "inherently unreliable, or an immediate bar to admissibility." *Harris*, 502 F. Supp. 3d at 42; *see also United States v. Simmons*, No. 2:16-cr-130, 2018 WL 1882827, at *5 (E.D. Va. 2018) ("The Court finds that all technical fields which require the testimony of expert witnesses engender some degree of subjectivity . . . which is based on specialized training, education, and relevant work experience."). Nor does a methodology's heavy reliance on the examiner's experience render it inadmissible. *See Kumho Tire*, 526 U.S. at 151 (contemplating "experienced-based methodology"). The bottom line is that many fields of science—like a doctor evaluating certain symptoms and using his or her professional judgment and experience to diagnose the patient—depend, to some extent, on an individual's qualitative judgment. *See Aman*, 748 F. Supp. 2d at 541 n.15. Nevertheless, because a considerable level of subjectivity influences any given firearm-toolmark examiner's conclusion, this factor weighs against admissibility.

v. General Acceptance

The last *Daubert* factor asks whether the methodology enjoys "general acceptance" in the expert's "relevant scientific community." *Daubert*, 509 U.S. at 594. In discussing the contours of that factor, the *Daubert* Court stated that widespread acceptance within a scientific community will weigh in favor of admissibility, while evidence based on a methodology "which has been able to attract only minimal support within the community may properly be

viewed with skepticism." *Id.* (quoting *United States v. Downing*, 753 F.2d 1224, 1238 (3d Cir. 1985)) (cleaned up).

Graham concedes that if the "relevant community [is] ballistics examiners, then this factor would weigh in favor of admissibility." (Def.'s Mot. Exclude at 24.) But he asserts that, because firearm-toolmark analysis is not a science and toolmark examiners are biased towards accepting the AFTE methodology, the court should look instead to a wider scientific community that has "generally . . . condemned [the discipline] for the current state of research on the validity of its methodology."[16] (*Id.*)

In support of a wider but undefined relevant community, Graham levels a familiar attack: "[T]he community of ballistics examiners has a vested career-based interest in the AFTE theory being accepted," so its general acceptance is unpersuasive. (Def.'s Mot. Exclude at 24 (citing *United States v. Shipp*, 422 F. Supp. 3d 762, 782 (E.D.N.Y. 2019).) Several courts view that line of reasoning as persuasive. *See, e.g.*, *Tibbs*, 2019 WL 4359486, at *21; *Briscoe*, 2023 WL 8096886, at *11; *Shipp*, 422 F. Supp. 3d at 782–83. For example, the *Tibbs* court believed that "if *Daubert* . . . and Rule 702 are to have any meaning at all, courts must not confine the relevant scientific community to the specific group of practitioners dedicated to the validity of the theory—in other words, to those whose professional standing and financial livelihoods depend on the challenged discipline." *Tibbs*, 2019 WL 4359486, at *21. But the court disagrees with that rationale and finds that the relevant community here is firearm-toolmark examiners.[17]

[16] Graham does not cite any authority for this proposition, so the court assumes he is referring to the NRC and PCAST Reports that he cited earlier in his brief.

[17] The court is also unpersuaded by the common attack that experts who use a certain methodology cannot be trusted to regulate their field. Of course, people with a vested financial interest would like to see the basis for

As an initial matter, the Fourth Circuit tacitly rejected the *Tibbs* court's analysis that undergirds much of Graham's argument. *See Crisp*, 324 F.3d at 268–699. In *Crisp*, Judge Michael's dissent echoed *Tibbs* while analyzing the relevant community for purposes of latent fingerprint examination:

> The fingerprint examination community is certainly a proponent of the technique. That community's enthusiasm, however, must be subjected to objective scrutiny if *Daubert* is to have any meaning. One author asserts that "mainstream scientists, by and large, have ignored the question of whether individuals can be reliably identified through small, distorted latent fingerprint impressions." At least two forensic commentators have expressed concern about the lack of objective scientific research into the reliability of the technique.

Id. at 276 (Michael, J., dissenting) (internal citations omitted). The majority disagreed, noting the forensic discipline enjoyed "a strong general acceptance, not only in the expert community, but in the courts as well," even if it had "not attained the status of scientific law." *Id.* at 268. In so doing, the *Crisp* court rejected the appellant's argument "that, while fingerprint analysis has gained general acceptance among fingerprint examiners themselves, this factor should be discounted because . . . the relevant community 'is devoid of financially disinterested parties such as academics.'" *Id.*

Additionally, *Kumho Tire*, *Daubert*, and Rule 702's liberal approach to admitting expert evidence supports that the relevant community in the instant case is firearm-toolmark

---

their careers continue to be viable. As the government notes, however, in rebutting the same argument as applied to a different factor, "[t]he same could be said of nearly any professional or scientific field, but we do not jettison entire fields of study because it is possible to cynically ascribe the worst of motivations to every practitioner." (Gov't Resp. at 24–25 n.15.) Overall, the court's focus is on reliability, and it only needs to "determine [if Hollenbeck's] method is reliable, not that it is free of any possibility of bias." *Pete*, 2023 WL 4928523, at *6 (quoting *Adams v. Lab'y Corp. of Am.*, 760 F.3d 1322, 1333 (11th Cir. 2014)).

examiners. In *Kumho Tire*, the Supreme Court posited that "it will be appropriate for the trial judge to ask . . . whether [an engineer's methodology] is generally accepted in the relevant engineering community." 526 U.S. at 151. Similarly, the Court noted that the relevant community for an expert perfume tester would be other people in the field of perfume testing. *Id.* Those examples demonstrate that the relevant community is other practitioners in the pertinent field, not the overall scientific community. *See id.* ("It is to make certain that an expert . . . employs in the courtroom the same level of intellectual rigor that characterizes the *practice* of an expert in the *relevant* field.") (emphasis added). That conclusion is also supported by one of *Daubert*'s chief purposes: rejecting the "rigid 'general acceptance' requirement" set forth by *Frye v. United States*, 293 F. 1013 (D.C. Cir. 1923). *Daubert*, 509 U.S. at 588. The *Daubert* Court rejected *Frye*'s "austere" general-acceptance standard because it was "at odds with the 'liberal thrust' of the Federal Rules" of Evidence. *Daubert*, 509 U.S. at 588–89. Of course, the Supreme Court did not do away with *Frye*'s edict entirely; though general acceptance should not be the "exclusive test," it can still be an important factor in deciding admissibility. *Id.* at 589, 594.

When discussing this fifth factor, *Daubert* also cited favorably to *United States v. Downing*, 753 F.2d 1224 (3d Cir. 1985), and its thorough discussion of the general-acceptance aspect of the *Frye* test. *See Daubert*, 509 U.S. at 594 (citing *Downing*, 753 F.2d at 1238). *Downing* examines how courts manipulated the phrase "relevant community" to admit or exclude evidence by too narrowly or too broadly defining the parameters of the relevant scientific community. *See Downing*, 753 F.2d at 1236–37. Defining the relevant community in this case as the scientific community at-large would result in exactly that type of improper manipulation and exclusion of evidence. On the other hand, it would also be improper for the court to define the relevant

scientific community here too narrowly: for example, as only DFS toolmark examiners. *Cf. Nease*, 848 F.3d at 232 (discussing how the relevant community for this factor was the relevant engineering community, not just engineers at a specific company). Defining the relevant community as other firearm-toolmark examiners therefore strikes the correct balance while remaining consistent with *Daubert*'s—and the Federal Rules'—goal of liberalizing the admission of expert evidence. *See Daubert*, 509 U.S. at 588–89; *Crisp*, 324 F.3d at 268–69.

As such, the court agrees with the majority of courts that "[t]he AFTE method easily satisfies this final factor." *United States v. Hunt*, 464 F. Supp. 3d 1252, 1259 (W.D. Okla. 2020), *aff'd* 63 F.4th 1229 (10th Cir. 2023). Toolmark identification enjoys widespread acceptance around the world. (*See* Gov't Resp. at 33 (citing *Harris*, 502 F. Supp. 3d at 42).) And acceptance of the AFTE methodology is undoubtedly widespread among the relevant community, as it is "the field's established standard." *Ashburn*, 88 F. Supp. 3d at 246. Moreover, the court doubts that the NRC and PCAST Reports constitute enough scrutiny so as to constitute general condemnation of the AFTE methodology.[18] (*See* Def.'s Mot. Exclude at 24.)

The last factor weighs in favor of reliability.

vi. *United States v. Crisp* is Persuasive

As previewed in the previous section, the court's *Daubert* analysis is buttressed by *Crisp*, 324 F.3d 261, in which the Fourth Circuit rejected a near identical challenge to the reliability of a similar forensic-comparison discipline under *Daubert*.

[18] As the government notes, the PCAST Report's authors themselves may not be fairly included within the relevant scientific community. (*See* Gov't Resp. at 9 (citing James Agar, II, *The Admissibility of Firearms and Toolmarks Expert Testimony in the Shadow of PCAST*, 74 Baylor L. Rev. 93, 129–31 (2022)); *Harris*, 502 F. Supp. 3d at 42–43.

The appellant in *Crisp* argued "the premises underlying fingerprinting evidence have not been adequately tested[,] . . . there is no known rate of error[,] fingerprint examiners operate without a uniform threshold of certainty required for a positive identification, and that fingerprint evidence has not achieved general acceptance in the relevant scientific community." *Crisp*, 324 F.3d at 266. The Fourth Circuit disagreed on each point, concluding that "[w]hile [appellant] may be correct that further research, more searching scholarly review, and the development of even more consistent professional standards is desirable, he has offered us no reason to reject outright a form of evidence that has so ably withstood the test of time." *Id.* at 269.

Although *Crisp* is readily distinguishable from the instant action, the Fourth Circuit's reasoning in dismissing a similar argument in a similar discipline is persuasive. The *Crisp* court affirmed admission of fingerprint identification, and, as the court clarified with Hollenbeck, toolmarks are essentially "fingerprints for guns." (Hr'g Tr. 17:4–5.) "[W]hile further research into [firearm-toolmark] analysis would be welcome, to postpone present in-court utilization of this bedrock forensic identifier pending such research would be to make the best the enemy of the good." *Crisp*, 324 F.3d at 270 (cleaned up).

vii. *Daubert* Conclusion

The court's gatekeeping duty "is a 'flexible one,' and it exercises 'broad discretion' in choosing which *Daubert* factors to apply and how to consider them." *Belville v. Ford Motor Co.*, 919 F.3d 224, 233 (4th Cir. 2019) (quoting *Oglesby v. Gen. Motors Corp.*, 190 F.3d 244, 250 (4th Cir. 1999)). Throughout his motion, Graham encourages the court to follow the approach taken by a select few courts in engaging in their own statistical and scientific analysis in

scrutinizing the discipline. Though that *may* be within a court's broad discretion—and some concerns with firearm-toolmark analysis are well-founded—the court concludes that such an active approach would be inconsistent with *Daubert*'s command. *See Daubert*, 509 U.S. at 600–01 (Rehnquist, C.J., concurring in part and dissenting in part) (noting that Rule 702 does not "impose[] on [courts] either the obligation or the authority to become amateur scientists in order to perform" their gatekeeping role).

On balance, four of the five *Daubert* factors weigh in favor of reliability here; the court therefore finds the methodology underlying Hollenbeck's opinion reliable.

2. Hollenbeck's Opinion Passes Muster Under Federal Rule of Evidence 702(d)

In addition to showing that the methodology underlying Hollenbeck's opinion is reliable, the government must also show that Hollenbeck's opinion "reflects a reliable application of the principles and methods to the facts of the case." Fed. R. Evid. 702(d). Though Graham's motion does not explicitly argue that Hollenbeck's opinion fails Rule 702(d), it implies as much. As Graham correctly notes, the 2023 amendments to Rule 702 require the court to take a closer look at Hollenbeck's conclusion than it previously would have. Nevertheless, the court finds that her conclusion reflects a reasonable application of DFS firearm-toolmark analysis methodology.

i. 2023 Amendments to Rule 702(d)

The 2023 amendments clarified that, under Rule 702(d), courts must examine an expert's conclusion to ensure that it follows from the methodology the expert relied on. Before the amendments, courts examined the methodology that an expert relied on but were reticent to analyze the conclusion itself. *See, e.g.*, *Bresler*, 855 F.3d at 195 ("[C]ourts may not evaluate

the expert witness' conclusion itself, but only the opinion's underlying methodology."). According to the advisory committee, such analysis was "an incorrect application of Rules 702 and 104(a)." Fed. R. Evid. 702 advisory committee's note to 2023 amendments. As of December 2023, Rule 702(d) enlisted courts to take a more active role in analyzing the expert's conclusion. *See id.* ("Rule 702(d) has also been amended to emphasize that each expert opinion must stay within the bounds of what can be concluded from a reliable application of the expert's basis and methodology.").

These amendments require an interesting balancing act. The court must analyze an expert's conclusion to ensure that it is the result of a reliable application of the methodology. *See* Fed. R. Evid. 702(d). But the court still must not overstep and become an "amateur scientist[]" in performing its gatekeeping role. *Daubert*, 509 U.S. at 600–01 (Rehnquist, C.J., concurring in part and dissenting in part). To strike this balance, the court must ask whether Hollenbeck's conclusions could be reasonably drawn from a reliable application of the DFS methodology; the court does not attempt to perform its own toolmark analysis or draw its own conclusions about if the tested shell casings likely came from the same firearm. *Cf.* Fed. R. Evid. 702 advisory committee's note to 2023 amendments ("Expert opinion testimony regarding the weight of feature comparison evidence (*i.e.*, evidence that a set of features corresponds between two examined items) must be limited to those inferences that can reasonably be drawn from a reliable application of the principles and methods.").

Rule 702(d) is therefore satisfied here if the government shows, by a preponderance of the evidence, that Hollenbeck's conclusion is a reasonable inference that is the result of a

reliable application of the DFS's toolmark-analysis methodology. Whether Hollenbeck's conclusion is irrefutably correct is not in the province of the court's gatekeeping duty.

ii. Hollenbeck's Application and Conclusion

Graham argues that Hollenbeck's examination is unreliable insofar as "there is no description—anywhere, not even in her testimony—about what parts of the toolmarks helped her to reach her conclusion. As a result, her opinion is impossible to verify, and her method is impossible to reproduce." (Def.'s Mot. Exclude at 6.) Indeed, Graham claims that "nothing in her report would allow any other examiner to determine why she reached her conclusion other than that her subject[ive] assessment of the pattern meant the shell casings shared a common origin." (*Id.*) At the outset, the court notes that these criticisms ring more as an indictment of the field and, for the reasons stated above, are unpersuasive. But to the extent Graham argues that Hollenbeck did not apply DFS's methodology reliably to reach a reasonable conclusion, the court disagrees and finds Hollenbeck's conclusion reliable under Federal Rule of Evidence 702(d).

Hollenbeck's testimony detailed her process as she conducted her examination. She first ran the samples through NIBIN, which identified that the shell casing recovered at the convenience store was a "potential association" to the test-fired cartridge cases.[19] (*See* Hr'g Tr. 27:1–6; Certificate of Analysis, Mar. 1, 2023 [ECF No. 55-3].) After Hollenbeck confirmed NIBIN's potential association and that the samples shared class characteristics, she used her

[19] Though both samples were 9mm Luger cartridge cases, they were produced by different manufacturers—the recovered-sample from CBC and the test-fired samples from Remington. (Hr'g Tr. 24:1–6.) Hollenbeck concedes that it would have been better to use the same brand of ammunition. (*Id.* 56:14.) But Graham does not argue that this defeats reliability, and the court thinks the same.

comparison microscope to compare the individual characteristics. (Hr'g Tr. 28:12–20.) In so doing, Hollenbeck considered "what type of breechface impression [she was] looking at, what type of firing [pin] impression, and . . . what the overall spatial relationship and pattern" of the individual characteristics said about the source of the toolmarks. (*Id.* at 28:17–29:5.) After her individual characteristic analysis, she concluded that the samples were, in her opinion, fired from the same gun.[20] (Hr'g Tr. 29:12–15.) As required by DFS procedure, Hollenbeck's report contains some notes, photographs, and standard language as to her conclusion to support her identification. (*See* Certificate of Analysis, May 31, 2023 [ECF No. 55-4]; DFS Manual at 23, 30, 59 (documentation standards), 76 (standard language).)

A second examiner came to the same conclusion after comparing the samples using the same methodology, further bolstering the notion that Hollenbeck's conclusion was a result of a reasonable application of the methodology. (Certificate of Analysis, May 31, 2023, at 4); *Harris*, 502 F. Supp. 3d at 43. While it may be true that not every toolmark examiner would have concluded that the marks warranted an 'identification' conclusion, the court is satisfied that it is more likely than not that Hollenbeck reliably applied the DFS's methodology and came to a reasonable conclusion.

In sum, Rule 702 still does not ask the court to "nitpick [Hollenbeck's] opinion in order to reach a perfect expression of what the basis and methodology can support," so long as she does not "make claims that are unsupported by [her] basis and methodology." Fed. R. Evid. 702 advisory committee's note to 2023 amendments. Any further skepticism about

[20] Hollenbeck testified that she looked for subclass characteristics during the microscopic comparison as well but found none, though she did not make note of that in her report. (Hr'g Tr. 29:22–30:3.)

Hollenbeck's conclusion is, as Graham's counsel effectively demonstrated at the hearing, "proper fodder not for the outright exclusion of evidence on *Daubert* grounds, but rather for robust cross-examination at trial." *Rhodes*, 2023 WL 196174, at *7; *see also Crisp*, 324 F.3d at 271 ("To the extent that a given [toolmark] analysis is flawed or flimsy, an able defense lawyer will bring that fact to the jury's attention, both through skillful cross-examination and by presenting expert testimony of his own.").

C. Limitations on Hollenbeck's Testimony

As an alternative to excluding Hollenbeck's testimony, Graham asks the court to limit it to allow her only "to testify to the fact that, in her mind, the patterns of toolmarks on the two shell casings appear similar." (Def.'s Mot. Exclude at 27.) The court agrees that some limitations are necessary, consistent with the specific concerns in the recent amendments to Rule 702 and those shrouding the discipline generally. *See Davis*, 2019 WL 4306971, at *6. But the court will not go as far as Graham asks. *Accord Romero-Lobato*, 379 F. Supp. 3d at 1117 (noting that among the courts that limit testimony, "the general consensus is that firearm examiners should not testify that their conclusions are infallible or not subject to any rate of error, nor should they arbitrarily give a statistical probability for the accuracy of their conclusions").

Consistent with the government's proposed stipulation, the court will limit Hollenbeck's testimony insofar as it must comply with the standard U.S. Department of Justice's Uniform Language for Testimony of Reports for the Forensic Firearms/Toolmarks Discipline ("DOJ ULTR") language. (Gov't Resp. at 41–42.) As relevant here, the DOJ ULTR defines a "source identification" as "an examiner's opinion that all observed class

characteristics are in agreement and the quality and quantity of corresponding individual characteristics is such that the examiner would not expect to find that same combination of individual characteristics repeated in another source and has found insufficient disagreement of individual characteristics to conclude they originated from different sources." (Gov't Resp. Ex. H at 2 [ECF No. 71-8].) An examiner who is called to testify regarding her source identification conclusion shall not assert that: "two toolmarks originated from the same source to the exclusion of all other sources," "examinations conducted in the forensic firearms/toolmarks discipline are infallible or have a zero error rate," or "two toolmarks originated from the same source with absolute or 100% certainty," among other restrictions. (*Id.* at 3.) The court thinks that DOJ ULTR's guidelines adequately address the concern that the jury will take Hollenbeck's word for absolute truth.[21] *See Harris*, 502 F. Supp. 3d at 45; *Hunt*, 464 F. Supp. 3d at 1261.

III. CONCLUSION

Graham "today advocates the wholesale exclusion of a long-accepted form of expert evidence. Such a drastic step is not required of [the court] under *Daubert*, and [the court] decline[s] to take it." *Crisp*, 324 F.3d at 268. At bottom, the court's gatekeeping duty is meant "to protect the judicial process from 'the potential pitfalls of junk science.'" *Sardis*, 10 F.4th at 275 (quoting *United States v. Bonner*, 648 F.3d 209, 215 (4th Cir. 2011)). Despite the valid criticisms of toolmark analysis, the government has shown that it is more likely than not that firearm-toolmark identification is not junk science. And the government met its burden of

[21] To the extent there is any ambiguity, the court makes clear that Hollenbeck may not testify to a "reasonable degree of scientific certainty" or similar expressions, which the DOJ's instructions agree should not be permitted. (Gov't Resp. Ex. H at 3.)

demonstrating that the DFS's firearm-toolmark identification methodology is, and Hollenbeck's application of that methodology was, reliable under *Daubert* and Federal Rule of Evidence 702(d). Accordingly, the court will permit Hollenbeck's testimony, consistent with the limitations set forth in this Memorandum Opinion.

The clerk is directed to forward a copy of this Memorandum Opinion and the accompanying Order to all counsel of record.

ENTERED this 20th day of February, 2024.

/s/ Thomas T. Cullen
HON. THOMAS T. CULLEN
UNITED STATES DISTRICT JUDGE